DOCUMENT RESUME

ED 152 841 TM 007 056

AUTHOR	Robertson, David W.; Montague, William E.
TITLE	Comparative Racial Analysis of Enlisted Advancement Exams: Relative Item-Difficulty Between Performance-Matched Groups. Final Report (August 1974- December 1975).
INSTITUTION	Navy Personnel Research and Development Center, San Diego, Calif.
REPORT NO	NPRDC-TR-76-34
PUB DATE	Apr 76
NOTE	29p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	Blacks; Caucasians; *Comparative Analysis; *Complexity Level; Enlisted Personnel; *Item Analysis; Matched Groups; Occupational Tests; Post Secondary Education; Promotion (Occupational); *Racial Differences; *Racial Discrimination; Tables (Data); *Test Bias; Test Items; Test Validity
IDENTIFIERS	Navy; *Navy Enlisted Advancement Examinations

ABSTRACT

In order to look for evidence of possible racial bias in six of the Navy tests for the Enlisted Advancement System, statistical data for a group of black personnel were compared with data for a selected matching group of whites. The first step was to compute item difficulty indices (the percent passing each item) for the black group, for the matching white group, and for a separate group of white personnel that were not matched with either group. A correlation was then computed for each test, between the item difficulty indices for the black group and the item difficulty indices for the nonmatched white group. Similarly, a correlation was computed for each test, between the item difficulty indices for the matched white group and the item and the item difficulty indices for the nonmatched white group. This experiment was repeated separately for four different pay grade levels, providing a total of 24 comparisons. In all comparisons, more similarity was found between the two white groups than between the black group and the nonmatched white group, although some differences were small. It was concluded that the tests may have some across-race bias, but no changes in test construction procedures were recommended at the time. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED152841

TM007 056

AFHRL-TR-76-34 APRIL 1976

COMPARATIVE RACIAL MEANS OF ENLIST[illegible] ADVANCEMENT EXAM RELATIVE ITEM-DIFFICULTY ESTIMATES PERFORMANCE-MATCHED GROUPS

David W. Zimmerman
William E. Sabers

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

76

2

NPRDC TR 76-34 April 1976

COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: RELATIVE ITEM-DIFFICULTY BETWEEN PERFORMANCE-MATCHED GROUPS

David W. Robertson
William E. Montague

Reviewed by
Robert P. Thorpe

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPRDC TR 76-34	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: RELATIVE ITEM-DIFFICULTY BETWEEN PERFORMANCE-MATCHED GROUPS	5. TYPE OF REPORT & PERIOD COVERED: Final Report (August 1974 - December 1975)
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): David W. Robertson, William E. Montague	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62763N, ZF55-521-031
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE: April 1976
	13. NUMBER OF PAGES: 27
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Item-analysis	Racial comparison
Test bias	Equal opportunity
Relative item-difficulty	Linear programming
Promotion	Matched groups

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Relationship between item-difficulty of racial groups was measured by the Rho value--the correlation between the two rank-orders of the Black and White item-difficulty levels. The notion of test bias in the Rho value is that, if the test is measuring the same type of ability in two groups, the same items should be relatively the easiest and most difficult for both groups--this should hold even though substantial differences may exist between two groups in total test scores or in percentage answering an item correctly. Since other studies have found that relative item-difficulty is a function of test-taking ability, this study

20. ABSTRACT (continued)

compared item-response relationships between Black and White groups matched on test score.

A "minority" matched White group was extracted from the total White group by applying a linear programming solution to minimize the squared difference between test scores of individuals in the Black group and the "minority" White group selected. Rho values were calculated: (1) across-race--between the Black and majority White groups; and (2) within-race--between the matched White and "majority" White groups. In the comparison of Rho values, lower across-race relationship than within-r ce relationship would be suggestive of test bias.

The linear programming procedure was quite useful in identifying test performance-matched groups. The across-race Rho values were slightly but consistently lower than those of within-race, suggesting that some small amount of bias (as defined for this analysis) may be present. The findings were considered only suggestive and tentative, since the analysis employed internal and conceptual, rather than external, criteria of test bias.

FOREWORD

This is the second of a series of studies which were initiated in response to a request from the Chief of Naval Personnel (Pers-6) to investigate the relative item-difficulty levels in enlisted advancement exams between Black and White racial groups, and to determine the feasibility of developing exams from items similar in difficulty for both racial groups as an approach to improving equal opportunity in career growth for minority groups. The first report in the series is NPRDC TR 76-6 "Comparative racial analysis of Enlisted Advancement Exams: Item-difficulty," July 1975. (Another study nearing completion concerns item-differentiation.)

The substantial and valuable assistance of the following persons is gratefully acknowledged: Mr. John G. Balaban, Ms. Belinda Halter, Mr. Greg C. Pascoe, and DP2 Suzanne Olson, for data processing and computation; and Ms. Hazel F. Schwab, for clerical support.

This study was performed under Exploratory Development Task Area ZF55-521-031 (Career Performance and Selection).

J. J. CLARKIN
Commanding Officer

SUMMARY

Problem and Purpose

A previous item-analysis of enlisted advancement exams investigated various test characteristics which might account for differences between Black and White racial groups. It was found that the relationship between item-difficulty for Blacks and Whites was low in some exams, as indicated by a low correlation, Rho value, between the two rank-orders of the Black and White item-difficulty levels. The finding suggested a possibility of racial bias in some of the exams. The Rho value is an "internal" type indicator of bias. The notion of test bias in this statistic is that, if the test is unidimensional and measures the same ability in both groups, the same items should be relatively the easiest and most difficult for both groups--this should hold even though substantial differences may exist between two groups in total test scores or in percentage answering an item correctly. It has been found that relative item-difficulty is a function of performance level on total score in studies of widely used selection tests of general intelligence. In contrast, the content of the tests of present interest is highly specific to Navy occupational skills. This study was performed to compare item-response relationships between Black and White racial groups matched on test score.

Approach

A "minority" matched White group, matched on Black test score, was extracted from the total White group by applying the Ford-Fulkerson linear programming procedure. Rho values were calculated: (1) across-race--between the Black and majority White groups; and (2) within-race--between the "minority" matched White and majority White groups. The Rho values were compared to determine if the across-race relationships were lower than the within-race relationships.

Findings and Conclusions

The use of the linear programming procedure selected closely matched racial groups (pages 9, 17). The distribution statistics of the residual White majority group changed very little (pages 9, 17). The across-race Rho values were slightly but consistently lower than within-race values (pages 12, 17). Thus, test performance on total test score appears to account largely, but not completely, for racial differences in relative item-difficulty. Since test bias was defined, for the present analysis, as across-race correlation less than within-race, the findings suggest that some small amount of bias may be present. However, it was emphasized that the findings should be considered only suggestive and tentative, since the analysis employed internal and conceptual, rather than external, criteria of test bias. (pages 17, 18)

Recommendation

No changes in test construction procedures are recommended at this time. Any such recommendations should be based upon analyses of item-differentiation (a presently ongoing study) and of item content based on external criteria (i.e., performance on the job to which promoted). (pages 17, 18)

CONTENTS

	Page
INTRODUCTION	1
Problem	1
Background	1
Purpose	4
METHOD	5
Data	5
Matching Procedure	5
Analysis	5
RESULTS	9
Precision of the Matching Procedure	9
Changes in the Residual White Majority	9
Differences in Relative Item-Difficulty	12
DISCUSSION AND FINDINGS	17
CONCLUSIONS AND RECOMMENDATIONS	18
REFERENCES	19

LIST OF TABLES

1.	Navy Advancement Exam Sample Sizes, Means and Standard Deviations by Racial Source Groups Prior to Matching Procedure	6
2.	Advancement Exam Sample Sizes, Means, and Standard Deviations by Racial Groups after Matching Procedure	10
3.	Item P Values by Racial Groups	11
4.	Relative Item-Difficulty (Rho Value) Between Racial Groups	13
5.	Relative Item-Difficulty (r Value) Between Racial Groups	14
6.	Significance Test of Differences in Relative Item-Difficulty	15

LIST OF TABLES (CONT.)

Page

7. Differences in Proportion of Accountable Variance Between Across-Race and Within-Race Correlations 16

FIGURE

1. Comparison of relative item-difficulty among pairs of racial groups 8

INTRODUCTION

Problem

The Enlisted Advancement System is one of the Navy's major personnel selection systems being studied to identify and alleviate any condition which might be detrimental to equal opportunity in career growth for all individuals and groups. Since a number of studies have found that a greater proportion of Caucasian than non-Caucasian candidates is advanced in the military services, an item-analysis of enlisted advancement exams was conducted to identify test characteristics which might account for differences between Black and White racial groups (Robertson & Royle, 1975).

One of the findings of the Robertson and Royle study was that the relative item-difficulty between Blacks and Whites is low in some exams, as indicated by a low correlation, Rho-value, between the two rank-orders of the Black and White item-difficulty levels. The finding suggested a possibility of racial bias in some of the exams.

There are a number of problems associated with an investigation of test bias, including particularly the definition of and criteria used in defining "bias." Use of an external criterion requires acquisition of a follow-up measure of performance effectiveness in the activities or behaviors for which the personnel selection was made. If, in the comparison of two groups, the selector test predicts the criterion behavior equally well for both groups, the selector test would probably not be considered biased (assuming that the criterion measure itself is fair). A test would be considered particularly unfair if it underpredicts the performance of one of the groups on the criterion performance (Thorndike, 1971).

Internal criteria of bias become important when discussing the construct validity of a test and in assessing bias, even if external criteria indicate no bias is present (Jensen, 1973), or if an external criterion is not yet available. However, findings of test bias resulting from an analysis without using an external criterion should be regarded as only tentative--that is, as clues of possible bias, rather than actual test bias (Echternacht, 1974). A finding that two groups score differently on a selector or on a criterion does not necessarily of itself support a finding of test bias. For example, in the Robertson and Royle (1975) study, although average total test scores and P values of Whites were higher than those of Blacks, this of itself is not sufficient evidence to conclude that the advancement tests are biased.

Background

Item characteristics which are useful in the study of test bias are: (1) Item X Group (race) interaction in an analysis of variance (ANOVA),

and (2) relative item-difficulty--Rho value--between the groups. The measure of the item-difficulty level is the percentage of a group which answers the item correctly (in the present study hereafter referred to as P value, i.e., the lower the P value, the more difficult is the item). The Rho value is the correlation, between the two race groups, of the rank-orders of the P values. (Analogously, the r value is the correlation of the P values themselves.) The notion of test bias in these characteristics is that, if the test is measuring the same ability in two groups, the size of the difference in P values found between the groups should be consistent, and the same items should thus be relatively the easiest and most difficult for both groups--this should hold even though substantial differences may exist between two groups in total test scores or in the percentage answering an item correctly. Most of the studies of bias of this type have analyzed standardized, widely used tests.

In an analysis of responses to the verbal and math sections of the Preliminary Scholastic Aptitude Test (PSAT) administered to Blacks and Whites in integrated schools of large metropolitan areas, Cleary and Hilton (1968) found significant but small Item X Race interactions, accounting for .9 to 1.8 percent of the total variance. It appeared that what little race effect was present was contributed by a "floor effect" of a few items--a convergence of Black and White item response differences on the most difficult items (those on which both races have a great tendency to guess) (Cleary and Hilton, 1968; Stanley, 1969).

Cardall and Coffman (1964) correlated the item difficulties of the verbal and math sections of the Scholastic Aptitude Test (SAT) of groups from three areas: (1) The rural Midwest, (2) New York City, and (3) the Southeast (where only Blacks were registered). The correlations for the verbal items between groups (3) and (1) and groups (3) and (2) were much lower than for those between groups (1) and (2).

Angoff and Ford (1973) compared PSAT responses of (1) Black and White groups randomly selected, and (2) Black and White groups matched on "related ability." These groups were from a large urban area in the Southeast. (Math section items were analyzed for Black and White groups matched [although not completely] on verbal section scores; and verbal section items were analyzed for groups matched on math section scores.) They found the correlation of item difficulties within races to be very high--.978 between the two Black samples, and .987 between the two White samples. The correlations across races (i.e., between Black and White groups) ranged from the high .80s to the low .90s--from .859 to .948. The correlations for the matched Black and White groups fell between the within- and across-race correlations--from .923 to .959--suggesting that the performance factor (i.e., on the related ability) contributes substantially to Item X Race interaction.

Jensen (1973) found similar results when he compared responses of matched groups on another type of test--the Wonderlic Personnel Test (WPT) of general intelligence--which is used widely in business and industry for selection and placement. The Race X Item interactions among

three subgroups of Whites (less Spanish surnames) and Blacks who were job applicants, at a company in a large city, were compared. The subgroups were: (1) the total number of Blacks and Whites (N = 204 each), (2) Blacks and Whites which could be matched ("equated") on WPT total score (N = 127 each), and (3) "Pseudo race" groups of Whites matched on WPT score with other Whites with a 7-point score difference (i.e., the approximate White-Black mean difference in national normative data) (N = 98 each). Results of the three ANOVA's performed on the various groups showed that the proportions of total variance (omega squared X 100) attributable to the Race X Item interaction were: (1) 1.04 percent ($p < .001$), (2) 0.29 percent ($p < .10$), and (3) 0.94 percent ($p < .001$). It is interesting to observe that when the Black and White groups were matched on test performance (in subgroup 2), the Race X Item interaction was reduced to a very small and statistically nonsignificant amount. However, when a White group was matched to a second White group having Black normative data (in subgroup 3), a highly significant "race" factor was produced. Jensen also calculated the relative item-difficulty (i.e., correlation of item P values) for the total Black and White sample (subgroup 1) and for nationwide samples of Black and White job applicants. The within-race correlations were near unity--nationwide-Black with city-Black $r = .992$ and White with White $r = .995$. The across-race correlations dropped substantially--nationwide-Black with nationwide-White $r = .932$ (on all items) and .916 (on attempted items only), and city-Black with city-White $r = .956$ (on all items) and .796 (on attempted items).

The Navy promotion tests analyzed by Robertson and Royle (1975), as distinguished from the tests described above, are not widely used, and do not measure general intelligence. Rather, the technical content of each 150-item, 4-choice test is highly specific to a Navy occupational specialty, and the number of advancement candidates varies widely across the specialties. Also, the membership within each occupational group is highly homogeneous, since each candidate has demonstrated hands-on proficiency and has studied the technical manuals in an occupational specialty. Of the 24 occupational groups analyzed (6 specialties across 4 seniority grade levels), Black sample sizes ranged from 26 to 193 and White, from 400 to 1817. (Sample sizes represented 100 percent of the Black and White populations examined.) Rho values for the 24 groups ranged from .551 to .932. The size of the Rho value was found to be a function of Black sample size. The correlation of the rank of the 24 Rho values with the rank of each Black sample size was $r = .598$.

The notion of relative item-difficulty as an indication of test bias suggests, of course, that the smaller the correlation coefficient, the greater the evidence of bias. An internal criterion is needed to compare the "smaller" coefficients. The Jensen (1973) and Angoff and Ford (1973) studies demonstrated that although across-race relationships were smaller than those of within-race (i.e., using the within-race correlation as the criterion), the differences were largely attributable to performance level on total score. The Robertson and Royle (1975) study identified sample size as a factor which reduced the size of the coefficients.

Purpose

The purpose of the present study, as a follow-on to the Robertson and Royle findings, is to determine whether racial difference in test performance also influences the relative item-difficulty in enlisted advancement tests. Racial groups which were matched both on test performance and sample size were investigated. The questions specifically addressed, for each occupational group, were:

1. Whether the Black minority group (B) can be matched on test performance with a White subgroup extracted from the majority group (i.e., a "minority" matched White group--MW).

2. Whether the extraction of the matched White subgroup (MW) from the White majority group appreciably changes the distribution statistics of the residual White majority group (WJ).

3. Whether the relationship between item-difficulty for the minority Black and residual majority groups differs from that between the matched White and residual majority groups.

METHOD

Data

The present analysis used the data set of the Robertson and Royle (1975) study--item responses to Navy Enlisted Advancement Exams for six occupational specialties across four pay grade levels (i.e., 24 different exams). The six specialties (which in the Navy are termed "Ratings") are:

Aviation Machinist's Mate (Jet Engine Mechanic) (ADJ)
Boatswain's Mate (BM)
Boiler Technician (BT)
Commissaryman (CS)
Hospital Corpsman (HM)
Machinist's Mate (MM)

The Black and White groups were extracted from the total population of all race groups which took the exams (Series 61 in August 1972). Table 1 presents race group by sample sizes, total test means, and standard deviations.

Matching Procedure

A matched White group was extracted from the total White group by applying the Ford-Fulkerson (1956) algorithm to match the extracted White group with the Black group on the basis of total test score. This algorithm provides a solution to what is commonly known as the "transportation problem" in which objects are moved from m origins to n destinations given the mn costs or distances involved. In the present application, the cost matrix was an n x m matrix of the squared differences in test score between the m individuals in the Black sample and the n individuals in the White sample. Those in the White sample with whom the individuals in the Black sample were paired by the minimization procedure constituted the matched White sample.

Thus, three race groups were selected for analysis:

1. B - the original minority Black group,
2. MW - a "minority" matched White group, matched to Black test score, and
3. WJ - a residual "majority" White group of all Whites except the relatively small proportion of White cases extracted to create the above matched White group.

Analysis

Total test score means, standard deviations, item $\underline{P}$ values, and $\underline{P}$ value ranks were calculated for each of the three racial groups within each of the 24 rate groups. (The Navy term "rate" identifies a Navy "rating" at a particular pay grade.)

Table 1

Navy Advancement Exam Sample Sizes, Means and Standard Deviations by Racial Source Groups Prior to Matching Procedure

Competition to Pay Grade	Rate	Race								
		Black			White			Other [a]		
		N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
4	ADJ3	47	52.38	12.60	644	69.96	14.75	21	56.86	12.99
	BM3	83	58.07	9.38	1033	64.15	11.86	10	58.30	12.52
	BT3	33	61.76	13.37	831	73.77	16.68	7	58.14	10.75
	CS3	27	67.59	10.15	447	76.12	11.76	17	79.12	6.95
	HM3	104	68.00	11.17	1429	73.45	15.53	22	73.50	14.00
	MM3	58	62.48	12.26	1259	72.44	16.56	49	59.27	13.86
5	ADJ2	30	58.27	14.39	565	63.55	15.01	74	62.27	14.12
	BM2	74	60.12	11.70	569	63.43	10.56	16	58.44	11.18
	BT2	28	60.11	10.25	511	73.61	16.57	12	57.42	14.65
	CS2	47	64.00	11.41	412	69.01	10.66	97	75.71	13.11
	HM2	111	63.60	9.43	1391	70.27	13.40	35	71.83	9.72
	MM2	30	56.37	13.69	984	74.09	15.95	29	59.62	13.30
6	ADJ1	50	67.78	15.56	400	72.31	15.19	14	63.29	14.14
	BM1	115	66.33	11.18	502	69.66	11.49	17	60.00	11.79
	BT1	79	70.44	13.57	495	80.70	17.18	7	69.43	17.61
	CS1	127	68.27	12.22	661	72.04	11.78	66	78.26	14.59
	HM1	26	68.58	6.87	546	71.32	11.08	32	69.63	12.09
	MM1	62	62.44	11.26	774	75.39	14.04	20	68.50	14.71
7	ADJC	88	66.77	14.23	1014	70.07	14.50	10	66.10	16.89
	BMC	193	63.60	12.42	1103	65.75	10.87	28	62.54	10.11
	BTC	138	77.91	17.61	956	80.57	15.59	20	84.05	15.53
	CSC	165	63.01	14.24	771	65.58	13.92	68	68.41	15.65
	HMC	157	71.24	13.73	1817	70.75	13.02	50	71.28	13.28
	MMC	110	75.35	13.81	1547	78.73	13.63	52	80.64	16.21

[a]Other - All other minority groups except Black combined.

Measures of relative item-difficulty between all possible pairs of racial groups were obtained by Pearson product-moment correlations of: (1) P value ranks (yielding Rho values), and (2) the P values themselves (yielding r values).[1] Thus, correlations were obtained between the following three race group pairs: (1) Black and White majority ($\rho_{B \cdot WJ}$), (2) matched White and White majority ($\rho_{MW \cdot WJ}$), and (3) Black and matched White ($\rho_{B \cdot MW}$).[2]

The relative item-difficulty (Rho and r) values among the various rate groups were compared. The comparison of primary interest was between the coefficients for the Black and White majority groups and the matched White and White majority groups. The comparisons are illustrated in Figure 1. In the notion of test bias, as defined in the present analysis, employing the within-race relationship (ρ_2 in Figure 1) as the criterion, if the across-race relationship is less than that of the within-race (i.e., $\rho_1 < \rho_2$), the difference is suggestive of test bias. To estimate the significance of the difference, Hotelling's t_{d_r} test was applied (Guilford, 1965, p. 190).[3]

---

[1]Computation of P value ranks by the Pearson and Spearman formulas would yield identical results (given no tied ranks). However, computation by the Pearson formula of ranks and of the values themselves should yield different results (Walker and Lev, 1969, p. 238).

[2]Calculation of within-race relative item-difficulty between randomly drawn White samples was also considered. However, the calculation was deemed unnecessary, since other studies had well established the existence of very high within-race correlations (e.g., Angoff and Ford, 1973; Jensen, 1973). It was thus assumed that the within-race correlation between two randomly drawn White samples would be substantially greater than that between a White sample drawn randomly and another White sample matched to Black scores.

[3]The application of the Hotelling test in the present analysis is atypical and probably in a way not intended. Consider the typical application--given three measures (X_1, X_2, and X_3) administered to Sample 1 (S_1), does X_3 demonstrate a higher (or lower) relationship with X_1 than does X_2? By contrast, consider the present design--given one measure of P values (X_1) administered to three independent racial group samples (S_1, S_2, and S_3), does S_3 demonstrate a higher (or lower) relationship with S_1 than does S_2?

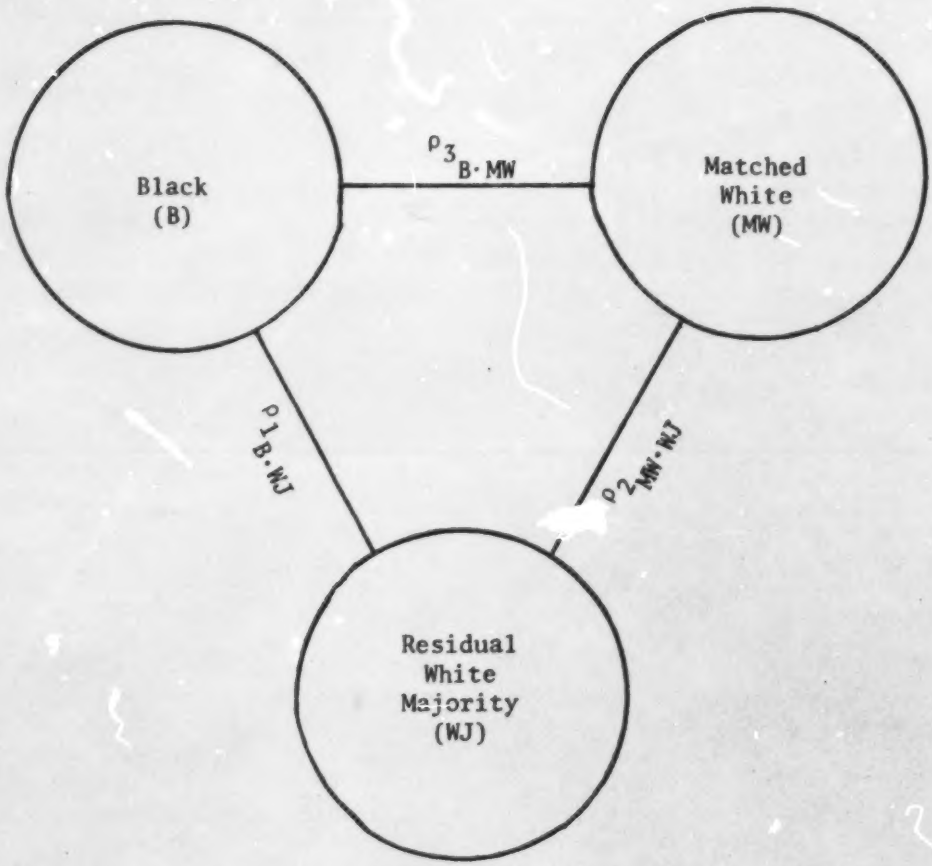

Figure 1. Comparison of relative item-difficulty among pairs of racial groups.

RESULTS

Precision of the Matching Procedure

The extremely disproportionate ratio of Blacks to Whites (see Table 1) made it possible: (1) to identify Whites with the same score as Blacks, and (2) to preserve the distribution characteristics of the White majority group from which the matched Whites were drawn. The application of the Ford-Fulkerson (1956) procedure was found to be highly effective in selecting individuals. Comparing the Total N and Identical Matched N columns of Table 2, all Black scores were perfectly matched in 6 of the 24 rate groups (e.g., BM3, all 83), all but 1 score in another 6 groups, and all but 2 or 3 scores in another 8 groups.

Comparing means and standard deviations, even in the rate groups with a number of nonperfect matches, the Black and matched White statistics are very similar (e.g., with 7 nonmatched ADJ3 scores, Black $\bar{X}$ = 52.38 and SD = 12.60, compared with matched White $\bar{X}$ = 52.36 and SD = 12.64).

P value ranges and medians of the Black and matched White groups are also very similar (see Table 3). (They are, of course, not quite identical, even in the perfectly matched groups, since the match was performed on total score, rather than on the P values themselves.)

Changes in the Residual White Majority

Since Black means were lower than White means before matching, it was reasonable to expect that the extraction of matched Whites would result in some increase in the residual White mean. Comparing Table 2 with Table 1, it may be observed that all residual White means (except HMC) increase in varying but still very small amounts. The amount of increase is, of course, a function of the original Black-White mean difference and, to a lesser extent, the variance and ratio of B:W sample size. Thus, the relatively large ADJ3 Black-White mean difference of 17.58 (in Table 1, 69.96 minus 52.38) results in a residual White mean of 71.34 (in Table 2), an increase of 1.38 points (about .10 σ); but the BM3 difference of 6.08 yields only a .53 increase (about .04 σ).[4]

---

[4]The difference between the original (W) and residual (WJ) White means were estimated in sigma unit differences ($\sigma_{diff.}$) by dividing the mean difference by the average standard deviation of the two groups, i.e.,

$$\sigma_{diff.} = \frac{\bar{X}_{WJ} - \bar{X}_{W}}{\frac{SD_{WJ} + SD_{W}}{2}}.$$ Thus,

for the BM3 example,

$$\frac{64.68 - 64.15}{\frac{11.91 + 11.86}{2}} = .04\sigma$$

Table 2

Advancement Exam Sample Sizes, Means, and Standard Deviations by Racial Groups After Matching Procedure

Competition to Rate	Racial Groups									
	Black (B)			Matched White (MW)				Residual White Majority (WJ)[b]		
	N	$\bar{X}$	SD	N	Ident. Match[a]	$\bar{X}$	SD	N	$\bar{X}$	SD
ADJ3	47	52.38	12.60	47	40	52.36	12.64	597	71.34	14.00
BM3	83	58.07	9.38	83	83	58.07	9.38	950	64.68	11.91
BT3	33	61.76	13.37	33	32	61.79	13.33	798	74.26	16.62
CS3	27	67.59	10.15	27	26	67.63	10.10	420	76.66	11.66
HM3	104	68.00	11.17	104	104	68.00	11.17	1325	73.88	15.74
MM3	58	62.48	12.26	58	58	62.48	12.26	1201	72.92	16.60
ADJ2	30	58.27	14.39	30	29	58.23	14.30	535	63.84	15.01
BM2	74	60.12	11.70	74	68	60.22	11.67	495	63.91	10.32
BT2	28	60.11	10.25	28	27	60.14	10.19	483	74.40	16.54
CS2	47	64.00	11.41	47	45	63.98	11.47	365	69.66	10.39
HM2	111	63.60	9.43	111	111	63.60	9.43	1280	70.85	13.54
MM2	30	56.37	13.69	30	28	56.67	13.14	954	74.64	15.73
ADJ1	50	67.78	15.56	50	47	67.56	14.84	350	72.98	15.14
BM1	115	66.33	11.18	115	112	66.31	11.13	387	70.66	11.43
BT1	79	70.44	13.57	79	75	70.46	13.54	416	82.65	17.11
CS1	127	68.27	12.22	127	125	68.28	12.18	534	72.93	11.52
HM1	26	68.58	6.87	26	26	68.58	6.87	520	71.46	11.23
MM1	62	62.44	11.26	62	60	62.47	11.20	712	76.51	13.70
ADJC	88	66.77	14.23	88	85	66.81	14.16	926	70.38	14.50
BMC	193	63.60	12.42	193	191	63.60	12.45	910	66.21	10.45
BTC	138	77.91	17.61	138	138	77.91	17.61	818	81.02	15.19
CSC	165	63.01	14.24	165	158	63.15	13.92	606	66.24	13.86
HMC	157	71.24	13.73	157	156	71.24	13.71	1660	70.71	12.95
MMC	110	75.35	13.81	110	109	75.34	13.82	1437	78.99	13.59

[a] N Whites matched to identical Black exam scores.

[b] The residual White majority (WJ) group contains all Whites except those matched with Blacks and extracted as the matched White (MW) group.

Table 3

Item P Values by Racial Groups

Competition to Rate	P Values					
	Black (B)		Matched White (MW)		White Majority (WJ)	
	Range	Mdn	Range	Mdn	Range	Mdn
ADJ3	8.5-78.7	34.0	8.5-61.7	34.0	9.9-85.1	46.7
BM3	8.4-89.2	38.7	6.0-86.7	39.8	8.2-90.0	44.0
BT3	6.1-90.9	42.4	9.1-87.9	39.4	6.9-94.7	49.0
CS3	3.7-96.3	44.4	0.0-96.3	44.4	3.1-96.4	49.3
HM3	5.8-95.2	45.2	5.8-98.1	45.2	7.0-96.2	50.0
MM3	6.9-82.8	39.7	8.6-82.8	39.7	9.1-88.9	48.6
ADJ2	10.0-73.3	36.7	6.7-76.7	40.0	11.2-75.3	42.9
BM2	6.8-81.1	39.3	4.1-81.1	40.6	5.1-85.5	43.0
BT2	3.6-85.7	39.3	0.0-82.1	39.3	11.4-85.3	50.7
CS2	2.1-91.5	42.6	2.1-95.7	40.4	2.5-95.9	45.2
HM2	2.7-97.3	41.4	2.7-96.4	41.4	4.1-96.1	47.0
MM2	3.3-83.3	36.7	0.0-90.0	36.7	7.4-87.7	49.9
ADJ1	12.0-92.0	44.0	8.0-94.0	46.0	9.1-92.9	48.4
BM1	8.7-90.4	42.6	6.1-87.8	43.5	5.7-93.8	46.8
BT1	8.9-89.9	45.6	8.9-93.7	45.7	8.7-95.9	56.7
CS1	4.7-83.5	45.8	5.5-82.7	47.3	8.1-88.2	51.2
HM1	0.0-100.0	42.3	0.0-96.2	42.3	2.7-98.1	45.3
MM1	4.8-93.5	40.3	8.1-90.3	40.3	11.4-97.6	49.4
ADJC	4.5-85.2	44.3	2.3-89.8	44.4	2.2-91.4	48.6
BMC	6.7-92.7	41.5	5.7-94.8	40.9	5.3-95.9	43.8
BTC	16.7-92.0	51.4	20.3-95.7	50.7	22.2-96.1	52.9
CSC	7.9-88.5	43.0	4.8-85.5	40.1	6.1-90.9	43.2
HMC	3.2-96.2	46.6	1.3-96.2	46.0	1.7-96.5	46.2
MMC	10.9-95.5	50.9	9.1-96.4	49.2	11.6-94.4	53.2

Changes in residual White standard deviations are very slight and variable.

Differences in Relative Item-Difficulty

The across-race (i.e., between the Black and White majority groups) correlations (i.e., the Rho values in Table 4 and r values in Table 5) were found to be lower than the within-race correlations (i.e., between the matched White and White majority groups). The differences are consistent in all 24 rate groups, as indicated by the B·WJ minus MW·WJ columns of Tables 4 and 5. The differences are small. However, to the extent that the application of the Hotelling test may be useful in estimating the significance of the difference, most of the differences are significant at the $p < .01$ level (see Table 6).

Horst (1966, p. 371) proposes that r^2 is more revealing than r for indicating the relative superiority, in terms of proportionate superiority, of one variable over another. Consider the square of the correlation values, multiplied by 100, as an indication of the percentage of the variance accounted for between racial groups (e.g., in Table 7, between the ADJ3 B·WJ groups, .506 x 100 = 50.6 percent). Although the differences between across-race and within-race values in Tables 4 and 5 appear small, the relative difference in proportion of accountable variance (right-hand column of Table 7) is substantial in some rate groups--at least 10 percent in 14 rate groups.

Table 4

Relative Item-Difficulty (Rho Value)
Between Racial Groups

Rate	Rho Value[a]			
	B·WJ[b]	MW·WJ[b]	B·WJ Minus MW·WJ	B·MW[b]
ADJ3	.710	.821	-.111	.753
BM3	.915	.951	-.036	.878
BT3	.778	.823	-.045	.638
CS3	.813	.884	-.071	.772
HM3	.946	.948	-.002	.917
MM3	.836	.913	-.077	.820
ADJ2	.586	.779	-.193	.397
BM2	.920	.939	-.019	.872
BT2	.707	.783	-.076	.614
CS2	.844	.908	-.064	.802
HM2	.879	.939	-.040	.879
MM2	.555	.716	-.161	.669
ADJ1	.773	.892	-.119	.657
BM1	.899	.952	-.053	.917
BT1	.862	.915	-.053	.871
CS1	.910	.951	-.041	.873
HM1	.871	.897	-.026	.841
MM1	.727	.875	-.148	.761
ADJC	.837	.932	-.095	.789
BMC	.907	.957	-.050	.890
BTC	.826	.943	-.117	.817
CSC	.900	.956	-.056	.877
HMC	.927	.965	-.038	.898
MMC	.845	.951	-.106	.813

Note. Racial Groups

B = Black
MW = Matched White
WJ = White Majority

[a]Correlation of P Value ranks.

[b]Rho $\geq$ .191 significant at $p < .01$ (1-tail), estimated by $t = \frac{\rho\sqrt{N - 2}}{\sqrt{1 - \rho^2}}$ (Hays, 1963, p. 646), average N = 147 items.

Table 5

Relative Item-Difficulty (r Value)
Between Racial Groups

Rate	r Value[a]			
	B·WJ[b]	MW·WJ[b]	B·WJ Minus MW·WJ	B·MW[b]
ADJ3	.711	.826	-.115	.746
BM3	.913	.947	-.034	.889
BT3	.798	.840	-.042	.684
CS3	.829	.893	-.064	.784
HM3	.952	.955	-.003	.931
MM3	.844	.914	-.070	.828
ADJ2	.628	.807	-.179	.428
BM2	.928	.946	-.018	.887
BT2	.718	.801	-.083	.642
CS2	.874	.917	-.043	.838
HM2	.924	.949	-.025	.912
MM2	.593	.739	-.146	.705
ADJ1	.809	.896	-.087	.694
BM1	.908	.958	-.050	.925
BT1	.883	.918	-.035	.893
CS1	.919	.962	-.043	.889
HM1	.881	.907	-.026	.855
MM1	.767	.903	-.136	.806
ADJC	.867	.943	-.076	.823
BMC	.930	.973	-.043	.916
BTC	.870	.953	-.083	.851
CSC	.915	.966	-.051	.902
HMC	.946	.975	-.029	.921
MMC	.872	.962	-.090	.840

Note. Racial Groups

B = Black
MW = Matched White
WJ = White Majority

[a]Correlation of P Values.

[b]$r \geq .191$ significant at $p < .01$ (1-tail), estimated by $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ (Hays, 1963, p. 529), average $N = 147$ items.

Table 6

Significance Test of Differences
In Relative Item-Difficulty

Rate	N Exam Items	Hotelling $t^a_{d_r}$	
		$\rho_{B.WJ} - \rho_{MW.WJ}$ [b]	$r_{B.WJ} - r_{MW.WJ}$ [c]
ADJ3	150	-3.46**	-3.59**
BM3	150	-3.40**	-3.11**
BT3	150	-1.38	-1.43
CS3	149	-3.03**	-2.95**
HM3	149	-0.23	-0.40
MM3	150	-4.11**	-3.86**
ADJ2	150	-3.88**	-4.05**
BM2	150	-1.66	-1.76
BT2	150	-1.90	-2.22*
CS2	149	-3.31**	-2.62**
HM2	149	-3.20**	-2.56*
MM2	150	-3.47**	-3.46**
AD[illegible]1	150	-4.61**	-3.75**
B[illegible]	150	-5.27**	-5.57**
BT1	150	-3.32**	-2.47*
CS1	148	-3.73**	-4.68**
HM1	150	-1.45	-1.59
MM1	149	-5.45**	-6.21**
ADJC	148	-5.46**	-5.27**
BMC	150	-4.90**	-6.07**
BTC	150	-7.36**	-6.55**
CSC	146	-5.12**	-5.79**
HMC	146	-4.50**	-4.72**
MMC	150	-7.41**	-7.83**

Note. Racial Groups

B = Black
MW = Matched White
WJ = White Majority

[a] df = N items minus 3. Negative t indicates B·WJ correlation is the smaller.

[b] Pearson coefficient applied to P value ranks.

[c] Pearson coefficient applied to P values.

*p < .05 two-tail test.

**p < .01 two-tail test.

Table 7

Differences in Proportion of Accountable Variance Between Across-Race and Within-Race Correlations

Rate	r^2			
	Across-Race B·WJ	Within-Race MW·WJ	B·WJ Minus MW·WJ	Proportionate Differences[a] %
ADJ3	.506	.682	-.176	-26
BM3	.834	.897	-.063	- 7
BT3	.636	.706	-.070	-10
CS3	.687	.797	-.110	-14
HM3	.907	.911	-.005	- 1
MM3	.712	.835	-.123	-15
ADJ2	.395	.650	-.256	-39
BM2	.860	.894	-.034	- 4
BT2	.515	.641	-.126	-20
CS2	.765	.840	-.076	- 9
HM2	.854	.901	-.048	- 5
MM2	.352	.546	-.194	-36
ADJ1	.654	.803	-.149	-19
BM1	.824	.919	-.095	-10
BT1	.779	.842	-.063	- 8
CS1	.844	.926	-.081	- 9
HM1	.776	.822	-.045	- 6
MM1	.588	.815	-.227	-28
ADJC	.751	.890	-.139	-16
BMC	.866	.946	-.080	- 8
BTC	.756	.907	-.151	-17
CSC	.838	.933	-.095	-10
HMC	.894	.950	-.056	- 6
MMC	.761	.925	-.165	-18

[a] Proportionate Difference = $\frac{r^2_{B\cdot WJ} - r^2_{MW\cdot WJ}}{r^2_{MW\cdot WJ}} \times 100$

DISCUSSION AND FINDINGS

When Whites were matched to Blacks on test performance, and each group was compared with the White majority group, three findings of primary interest resulted. They are presented in the order of the questions addressed in the Purpose section (page 4).

1. Since the proportion of White candidates competing for Navy enlisted advancement was much greater than Blacks, it was possible to extract "minority" groups of Whites which were identically or closely matched on Black total test score. The adaptation of the linear programming model was very useful to achieve an effective matching procedure.

2. The distribution statistics of the residual White majority group changed very little as a result of the extraction of the matched White groups (which, of course, was also due to the much greater proportion of White than Black candidates).

3. When measures of relative item-difficulty were compared between across-race groups (of White majority and minority Black) and within-race groups (of White majority and Whites matched to Black test score), both within- and across-race correlations were low--much lower than similar measures reported in the literature for verbal, math, and general intelligence types of tests used for selection purposes. For the present Navy tests, the across-race values were slightly but consistently (in all 24 Rate groups) and significantly (in 20 of 24 groups) lower than the within-race values.

Thus, test performance as measured by total test score appears to account for much, but not all, of the difference between across- and within-race values on tests of occupation-specific content (i.e., the present advancement tests), as well as on tests of general intelligence (e.g., WPT) and of scholastic verbal and math ability (e.g., PSAT). The implication of the above third finding is that a measure of relative item-difficulty may not be a useful indicator of bias if the two groups being compared differ on mean total-test score (i.e., as these Navy racial groups do--when no matching procedure has been applied). However, when matched groups are compared, and across-race correlations are found to be smaller than the within-race values, the types of items which account for the differences would be of interest.

It was beyond the scope of the present study to identify these types of items. In the "floor effect" from the most difficult items, as observed by Cleary and Hilton (1968) and Stanley (1969), the leveling-out of the most difficult Black $\underline{P}$ values may be more extended than for those of the matched White group. Further, since Black $\underline{P}$ values were more similar to White on applied than on theoretical-type item content, as observed by Robertson and Royle (1975) on some exams, the condition may tend to prevail even when Whites are matched to Blacks on total test score. Thus, it may be useful to investigate the most difficult items and applied versus theoretical-type items. However, analysis of item types is not deemed appropriate until an external criterion (i.e., subsequent performance on the job for which selected, in the present case, for which promoted) is employed.

CONCLUSIONS AND RECOMMENDATIONS

The conclusions resulting from the present study of relative item-difficulty in promotion exams of occupation-specific technical content are as follows:

1. Test performance on total test score appears to account largely, but not completely, for racial differences in relative item-difficulty.

2. Since test bias was defined, for the present analysis, as across-race correlation less than within-race correlation, and the across-race correlations were in fact found to be slightly but consistently less than within-race correlations, some small amount of bias may be present. (However, it is emphasized that the finding should be considered only suggestive and tentative, since the analysis employed internal and conceptual, rather than external, criteria of test bias.)

No changes in test construction procedures are recommended at this time. Any such recommendations should be based upon analyses of item-differentiation (a presently ongoing study) and of item content based on external criteria (i.e., performance on the job to which promoted).

REFERENCES

Angoff, W. H., & Ford, F. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Cardall, C., & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. Princeton, N.J.: Educational Testing Service, 1964, Research Bulletin 64-61. Cited by Cleary, T. A., & Hilton, T. L., An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Cleary, T. A. & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Ford,L. R., Jr., & Fulkerson, D. R. Solving the transportation problem. Management Science, 1956, 3, 24-32.

Guilford, J. P. Fundamental statistics in psychology and education (4th ed.). New York: McGraw-Hill, 1965.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Horst, P. Psychological measurement and prediction. Belmont, California: Wadsworth, 1966.

Jensen, A. R. An examination of culture bias in the Wonderlic Personnel Test. Paper presented at the National Academy of Sciences, Washington, D. C., October 22, 1973.

Robertson, D. W., & Royle, M. H. Comparative racial analysis of Enlisted Advancement Exams: Item-difficulty. San Diego: Navy Personnel Research and Development Center, July 1975, Technical Report TR 76-6.

Stanley, J. C. Plotting ANOVA interactions for ease of visual interpretation. Educational and Psychological Measurement, 1969, 29, 793-797.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Walker, H. M., & Lev, J. Elementary statistical methods (4th ed.). New York: Holt, Rinehart and Winston, 1969.

DISTRIBUTION LIST

Assistant Secretary of the Navy (Manpower and Reserve Affairs)
Office of Assistant Secretary of Defense (M&RA), Washington, D. C.
Chief of Naval Operations (OP-39) (OP-59) (OP-098T) (OP-098TL)(2)
(OP-099)(2) (OP-103B) (OP-914) (OP-964) (OP-987F) (OP-987P10)
Chief of Naval Personnel (Pers-1) (Pers-10c) (Pers-2B) (Pers-21)
(Pers-212) (Pers-2120) (Pers-5) (Pers-523) (Pers-6) (Pers-611)(2)
Chief of Naval Technical Training (Code 016) (Code N45)
Chief of Naval Education and Training Support
Chief of Information (OI-2252)
Chief of Naval Material (NMAT 0344) (NMAT 035)
Chief of Naval Research (Code 450)(4) (Code 458)(2)
Commander Navy Recruiting Command
Commander, Naval Electronics Laboratory Center (2)
Commanding Officer, Naval Education and Training Program Devel. Center (3)
Commanding Officer, Naval Health Research Center
Director, Behavioral Sciences Dept., Naval Medical Research Inst., Bethesda
Science Advisor, ACOS for Tactical Development, Commander Second Fleet
Superintendent, Naval Academy
Superintendent, Naval Postgraduate School
Superintendent, United States Military Academy
Superintendent, Air Force Academy
Superintendent, Coast Guard Academy
Technical Library, Naval Postgraduate School
Human Goals Office, Naval Education and Training Center, Newport, R. I.
Army Research Institute for Behavioral and Social Sciences
U. S. Army Enlisted Evaluation Center, Fort Benjamin Harrison (2)
Human Resources Development Division, U. S. Army Personnel and
Administration Combat Developments Activity
Chief of Research and Development, Department of the Army,
(Attn: Army Personnel Research Office), (Attn: Human Factors and
Operations Research Division), Washington, D. C.
Commander, Air Force Human Resources Laboratory, Brooks Air Force Base
Headquarters, Air Force (Attn: Science Division, Directorate of Science
and Technology, DCS/Research and Development), Washington, D. C.
Program Manager, Life Sciences Directorate, Air Force Office of
Scientific Research (AFSC)
Personnel Research Division, Air Force Human Resources Laboratory
(AFSC), Lackland Air Force Base
Occupational and Manpower Research Division, Air Force Human Resources
Laboratory (AFSC), Lackland Air Force Base
Technical Library, Air Force Human Resources Laboratory (AFSC),
Lackland Air Force Base
Defense Race Relations Institute, Patrick Air Force Base
Headquarters, Coast Guard, (Attn: Chief, Training and Procurement
Division), (Attn: Psychological Research Branch, (G-P-1/62),
Washington, D. C.
U. S. Coast Guard Institute, Oklahoma City, Oklahoma
Center for Naval Analyses, Arlington, Virginia
Director, Defense Documentation Center, Alexandria (12)
Director, Utilization and Management Manpower Techniques,
Washington, D. C.
National Research Council, Washington, D. C.
National Science Foundation, Washington, D. C.